MAINTENANCE METHOD OF RETRIEVAL SYSTEM AND RETREIVAL SYSTEM

Patent number: JP2001337973 (A)

Publication date: 2001-12-07

Inventor(s): MATSUI KAZUKI; IWAYAMA NOBORU; KIHARA HIDETO +

Applicant(s): FUJITSU LTD +

Classification:

- international: ***G06F12/00; G06F13/00; G06F17/30;*** **G06F12/00; G06F13/00; G06F17/30;** (IPC1-7): G06F12/00; G06F13/00; G06F17/30

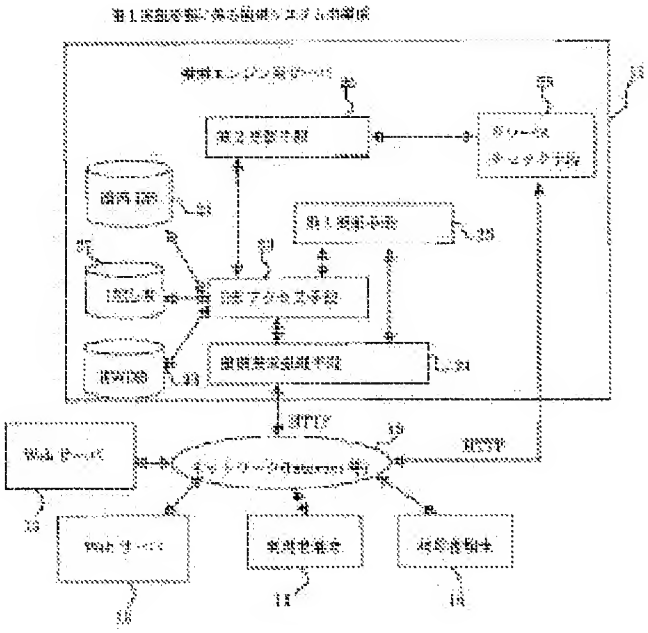
- european:

Application number: JP20000154502 20000525

Priority number(s): JP20000154502 20000525

Abstract of **JP 2001337973 (A)**

PROBLEM TO BE SOLVED: To enhance the retrieval precision in a retrieval system. SOLUTION: A retrieval DB is retrieved by use of a keyword ranked higher in a KWDB 22 to check the presence of link disconnection for the web page of the retrieval result. When a link disconnection occurs as the result of the check, the corresponding URL is eliminated from the retrieval DB 21. This processing is performed more frequently than other web pages. For a web page of the field conceivably attracting high users' attention, the link disconnection can be highly frequently checked, and the useless URL can be eliminated from the retrieval DB. Consequently, the retrieval precision can be efficiently enhanced.



(19)日本国特許庁（ＪＰ）　(12) **公開特許公報（Ａ）**　(11)特許出願公開番号
特開2001－337973
（P2001－337973A）
(43)公開日　平成13年12月７日(2001.12.7)

(51)Int.Cl.[7]		識別記号	FI		テーマコード(参考)
Ｇ０６Ｆ	17/30	２４０	Ｇ０６Ｆ　17/30	２４０Ａ	５Ｂ０７５
		１１０		１１０Ｆ	５Ｂ０８２
		１３０		１３０Ｂ	
		３４０		３４０Ｂ	
	12/00	５４６	12/00	５４６Ｂ	

審査請求　未請求　請求項の数８　ＯＬ　（全 13 頁）　最終頁に続く

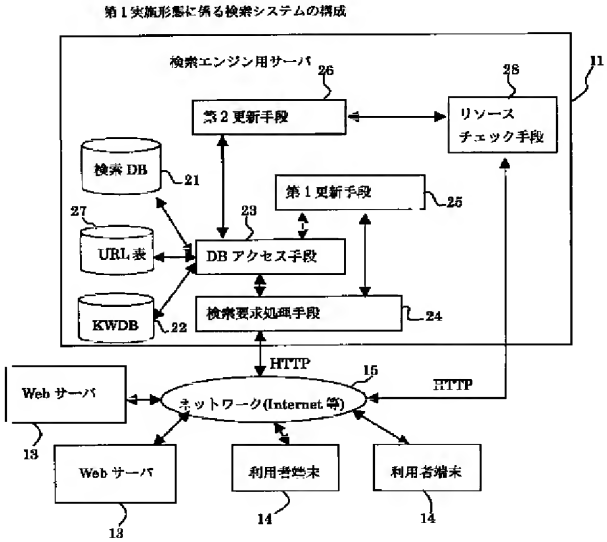
(21)出願番号　特願2000－154502(P2000－154502)
(22)出願日　平成12年５月25日(2000.5.25)

(71)出願人　000005223
富士通株式会社
神奈川県川崎市中原区上小田中４丁目１番１号
(72)発明者　松井　一樹
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内
(72)発明者　岩山　登
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内
(74)代理人　100094145
弁理士　小野　由己男　（外２名）

最終頁に続く

(54)【発明の名称】　検索システムのメンテナンス方法及び検索システム

(57)【要約】
【課題】　検索システムにおける検索精度を高める。
【解決手段】　ＫＷＤＢ２２において上位にランク付けされているキーワードを用いて検索ＤＢを検索し、検索結果のウェブページについてリンク切れの有無をチェックする。チェックの結果、リンク切れが生じていれば、該当ＵＲＬを検索ＤＢ２１から削除する。この処理を、他のウェブページよりも頻繁に行う。利用者に関心が高いと考えられる分野のウェブページについては、高い頻度でリンク切れをチェックでき、検索ＤＢから無駄なＵＲＬを排除できる。その結果、効率よく検索精度を高めることが出来る。



【特許請求の範囲】
【請求項１】ネットワーク上のリソースを利用者端末からの検索要求に応じて検索する検索システムのメンテナンス方法であって、
所定の基準で分類されたリソースを蓄積する検索蓄積手段と、リソースを検索するために用いられたキーワードを所定の順序でランク付けして蓄積するキーワード蓄積手段とを準備し、
リソースに順次アクセスしてリソースが無効になっているか否かをチェックし、かつキーワード蓄積手段で上位にランク付けされているキーワードにヒットするリソースについては前記チェックの頻度を高め、
リソースのチェック結果に基づいて検索蓄積手段を更新する、
検索システムのメンテナンス方法。
【請求項２】前記キーワード蓄積手段のキーワードは、検索に用いられた使用回数順にランク付けされている、請求項１に記載の検索システムのメンテナンス方法。。
【請求項３】前記キーワード蓄積手段のキーワードは、所定期間内におけるキーワードの使用回数の変化に基づいてランク付けされている、請求項１に記載の検索システムのメンテナンス方法。
【請求項４】禁止されている用語を蓄積した禁止用語蓄積手段をさらに準備し、
キーワード蓄積手段で上位にランク付けされているキーワードが禁止用語蓄積手段に蓄積されている用語に含まれる場合、当該キーワードについての前記チェックを行わない、請求項１～３のいずれかに記載の検索システムのメンテナンス方法。
【請求項５】所定の類似用語を同一の用語と見なしてキーワード蓄積手段を構築する、請求項１～４のいずれかに記載の検索システムのメンテナンス方法。
【請求項６】前記キーワード蓄積手段は、所定の分類基準に基づいて複数準備されている、請求項１～５のいずれかに記載の検索システムのメンテナンス方法。
【請求項７】ネットワーク上のリソースを利用者端末からの検索要求に応じて検索する検索システムであって、
所定の基準で分類されたリソースを蓄積する検索蓄積手段と、
検索に用いられたキーワードを所定の順序でランク付けして蓄積するキーワード蓄積手段と、
利用者端末からの検索要求に基づいて検索蓄積手段を検索し、検索結果を要求元に通知する要求処理手段と、
検索毎にキーワード蓄積手段を更新する第１更新手段と、
リソースにアクセスしてリソースが無効になっているか否かをチェックし、かつキーワード蓄積手段で上位にランク付けされているキーワードにヒットするリソースについては、前記チェックを優先的に行うリソースチェック手段と、
リソースのチェック結果に基づいて、検索蓄積手段を更新する第２更新手段と、
を備えた検索システム。
【請求項８】ネットワーク上のリソースを利用者端末からの検索要求に応じて検索する検索システムであって、
所定の基準で分類されたリソースを蓄積する検索蓄積手段と、
所定の分類基準ごとに設けられ、検索に用いられたキーワードを所定の順序でランク付けして蓄積する複数のキーワード蓄積手段と、
利用者端末からの検索要求に基づいて検索蓄積手段を検索し、検索結果を要求元に通知する要求処理手段と、
検索要求に基づいて更新すべきキーワード蓄積手段を特定し、特定したキーワード蓄積手段を検索毎に更新する更新手段と、
を備えた検索システム。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明は、インターネットなどで利用できるWWW(World Wide Web)の検索システムに関する。
【０００２】
【従来の技術】インターネット上には、各種の検索システムが存在している。これらの検索システムでは、キーワードなど所定の基準に基づいてウェブページを分類し、分類結果としてウェブページの存在場所を示すアドレス、例えばＵＲＬ(Uniform Resource Locator)を検索ＤＢ（データベース）に蓄積している。そして、この検索ＤＢに基づいて、入力されたキーワードにヒットするウェブページを検索する。
【０００３】表１は、検索ＤＢに蓄積される情報の概念を示す。一般的に、検索ＤＢへの蓄積の仕方として、ウェブページのＵＲＬと、そのウェブページに記述されている文章等のコンテンツに含まれる単語等のキーワード情報とを対応付けた表を作る。 そして、検索キーワードとともに検索要求が送られてくると、この表を検索してキーワードにマッチするURLを返す処理を行う。
【０００４】
【表１】

「キーワードに基づいて分類されたURL表」の例

URL	キーワード
http://www.fujitsu.co.jp	富士通
http://www.flab.fujitsu.co.jp	富士通,研究所
…	…

インターネット上の検索システムでは、通常このテーブルをインターネット上でアクセス可能な全てのウェブページについて作成する。よって、このテーブルは非常に膨大なデータ量に達するため、その維持・管理に多くの処理時間がかかる。

【０００５】具体的には、ウェブページは、他のＵＲＬに移動したり、消滅したりして変化するので、検索システムではその変化を検索ＤＢに反映させなければならない。そのため検索システムは、常にインターネット上のウェブページを巡回し、未登録のものを検索ＤＢに登録すると共に、その変化を検出して検索ＤＢを更新している。

【０００６】検索ＤＢの更新方法には、大別して２つの方法がある。１つは、メタ検索と呼ばれる方法である。この方法では、ある一つの検索エンジンが利用者の検索要求を受けると、複数の検索エンジンがその検索を実行する。そして、戻ってきた検索結果の中からリンク切れになっているものを探しだし、検索ＤＢから削除する方法である（MetaCrawler等）。

【０００７】他の１つの方法は、クライアント端末側でフィルタリング処理する方法である。すなわち、通常利用者の端末にインストールされている検索用ソフトウェアにより、検索結果の中でリンク切れのＵＲＬを削除してから表示するものである。

【０００８】

【発明が解決しようとする課題】しかし、昨今のインターネットの利用拡大に伴い、ウェブページ数が飛躍的に増大し、検索システムがウェブページの変化についていけなくなってきている。例えば、検索結果のＵＲＬにアクセスした結果、“Not Found（リンク切れ）”になったり、検索結果に同一のＵＲＬが何個もでたりする場合が多々ある。

【０００９】従来の検索システムでは、元々の検索ＤＢ２１の中にリンク切れのウェブページの情報が残っている。そのため、検索速度が遅くなるとともに、利用者にとって不要なウェブページが検索結果に残ってしまい、検索の精度を下げることになっている。これは、検索ＤＢの更新頻度が、ウェブページの更新頻度に比べて低いことに大きな原因がある。

【００１０】なぜなら、リンク切れをチェックするために、膨大な量のウェブページを一巡してそのウェブページの存在場所に変更がないかを確認しようとすると、現在の検索システムの能力では２週間以上要することが通常である。一方、ウェブページは、毎日のように更新されるものが少なくない。その結果、検索ＤＢには、実際のウェブページを反映していない情報が蓄積されたままとなり、検索結果の精度が下がってしまう。

【００１１】一方、いくつかの検索システムでは、検索キーワード（ＫＷ）のランキング情報を蓄積したキーワードＤＢを公開している。キーワードＤＢのランキング情報を見ると、一般的に検索キーワードは、検索に使用された時間が近い場合、類似する傾向にある。例えばその時に話題になっている製品の機種、ニュース、有名人の名前などが、検索キーワードのランキング上位になることが多い。しかし、検索システムにおいて、キーワードＤＢを活用して検索やその他の処理に用いた例はまだない。

【００１２】本発明は、前記課題を解決するために、無駄な検索結果が含まれることを防止し、ウェブページの検索精度を向上させることを目的とする。また、本発明は、キーワードＤＢを活用して検索システム上で提供されるサービスの質をさらに向上させることを目的とする。

【００１３】

【課題を解決するための手段】前記課題を解決するために、本願第１発明は、ネットワーク上のリソースを利用者端末からの検索要求に応じて検索する検索システムのメンテナンス方法であって、

Ａ；所定の基準で分類されたリソースを蓄積する検索蓄積手段と、リソースを検索するために用いられたキーワードを所定の順序でランク付けして蓄積するキーワード蓄積手段とを準備し、

Ｂ；リソースに順次アクセスしてリソースが無効になっているか否かをチェックし、かつキーワード蓄積手段で上位にランク付けされているキーワードにヒットするリソースについては前記チェックの頻度を高め、

Ｃ；リソースのチェック結果に基づいて検索蓄積手段を更新する、

検索システムのメンテナンス方法を提供する。

【００１４】ネットワーク上のリソースとは、例えばインターネット上のウェブページである。検索蓄積手段には、例えばキーワードに基づいて分類されたウェブページのＵＲＬが蓄積されている。キーワード蓄積手段には、ウェブページの検索に用いられたキーワードが、例えば使用回数順に蓄積されている。キーワード蓄積手段は、検索が行われるたびに更新される。

【００１５】この検索システムは、ネットワーク上のウェブページに次々にアクセスし、そのウェブページが削除されていたり他に移動してリンク切れになっているかどうかをチェックする。このチェックをインターネット上の全てのウェブページについて行うと、各ウェブページについてのチェック頻度が例えば２～４週間に１回程度になってしまう。そこで、検索頻度の多いウェブページ、すなわちよく使用されるキーワードを含むウェブページについては、チェックを高い頻度、例えば１日１回行う。リンク切れが生じていると判定された場合、そのＵＲＬは無効なので検索蓄積手段から削除される。

【００１６】インターネット上の全てのウェブページをもれなく同じ頻度でチェックするのではなく、使用頻度が高いと考えられるウェブページについてのみチェック

頻度を高くすることにより、ウェブページの検索速度と検索精度とを効率的に向上させる。本願第２発明は、前記第１発明において、前記キーワード蓄積手段のキーワードは、検索に用いられた使用回数順にランク付けされている検索システムのメンテナンス方法を提供する。

【００１７】利用者により検索される頻度が高いウェブページについては、リンク切れを頻繁にチェックすることが好ましい。そこで、キーワードを使用頻度順にランク付けし、使用頻度が高いキーワードを含むウェブページについてチェック頻度を高める。本願第３発明は、前記第１発明において、前記キーワード蓄積手段のキーワードは、所定期間内におけるキーワードの使用回数の変化に基づいてランク付けされている検索システムのメンテナンス方法を提供する。

【００１８】例えば、急激に使用回数が高まったキーワードは、利用者の関心の高まりを表していると考えられる。そのようなキーワードについては、総使用回数が多くなくても、そのキーワードを含むウェブページを頻繁にチェックすることが好ましいと考えられる。本願第４発明は、前記発明の方法において、禁止されている用語を蓄積した禁止用語蓄積手段をさらに準備し、キーワード蓄積手段で上位にランク付けされているキーワードが禁止用語蓄積手段に蓄積されている用語に含まれる場合、当該キーワードについての前記チェックを行わない検索システムのメンテナンス方法を提供する。

【００１９】例えば、差別用語や放送禁止用語などを、禁止用語蓄積手段に蓄積しておく。このような用語については、例えキーワード蓄積手段で上位になっても、リンク切れのチェックを行わない。本願第５発明は、前記発明の方法において、所定の類似用語を同一の用語と見なしてキーワード蓄積手段を構築する検索システムのメンテナンス方法を提供する。

【００２０】例えば既存の類似用語辞書を用いて類似用語群を同一キーワードとみなす。例えば、空白や長音の有無、平仮名やカタカナ、ローマ字の違いのみしかない用語を、同一のキーワードと見なし、使用回数を算出する。実情に即したキーワード蓄積手段の構築が可能となり、ひいては検索精度の向上を高めることが出来る。本願第６発明は、前記発明の方法において、前記キーワード蓄積手段が所定の分類基準に基づいて複数準備されている、検索システムのメンテナンス方法を提供する。例えば、ドメイン名や検索の時間帯ごとにキーワード蓄積手段を設けておく。

【００２１】本願第７発明は、ネットワーク上のリソースを利用者端末からの検索要求に応じて検索する検索システムであって、検索蓄積手段と、キーワード蓄積手段と、要求処理手段と、第１更新手段と、リソースチェック手段と、第２更新手段と、を備えた検索システムを提供する。検索蓄積手段は、所定の基準で分類されたリソースを蓄積する。キーワード蓄積手段は、検索に用いられたキーワードを所定の順序でランク付けして蓄積する。要求処理手段は、利用者端末からの検索要求に基づいて検索蓄積手段を検索し、検索結果を要求元に通知する。第１更新手段は、検索毎にキーワード蓄積手段を更新する。リソースチェック手段は、リソースにアクセスしてリソースが無効になっているか否かをチェックし、かつキーワード蓄積手段で上位にランク付けされているキーワードにヒットするリソースについては、前記チェックを優先的に行う。第２更新手段は、リソースのチェック結果に基づいて、検索蓄積手段を更新する。

【００２２】第１発明のメンテナンス方法を適用した検索システムである。本願第８発明は、ネットワーク上のリソースを利用者端末からの検索要求に応じて検索する検索システムであって、検索蓄積手段と、キーワード蓄積手段と、要求処理手段と、更新手段と、を備えた検索システムを提供する。検索蓄積手段は、所定の基準で分類されたリソースを蓄積する。キーワード蓄積手段は、所定の分類基準ごとに複数設けられ、検索に用いられたキーワードを所定の順序でランク付けして蓄積する。要求処理手段は、利用者端末からの検索要求に基づいて検索蓄積手段を検索し、検索結果を要求元に通知する。更新手段は、検索要求に基づいて更新すべきキーワード蓄積手段を特定し、特定したキーワード蓄積手段を検索毎に更新する。

【００２３】この検索システムでは、キーワード蓄積手段が、所定の分類ごとに複数設けられている。例えば、ドメイン名、利用者の年代や居住地域、利用時間帯などにより、キーワード蓄積手段を分類する。例えばドメイン名毎にキーワード蓄積手段を設けた場合、各キーワード蓄積手段は、該当するドメイン名を有する利用者端末から検索要求があったときに更新される。

【００２４】

【発明の実施の形態】＜発明の概要＞本発明は、インターネット上のウェブページ（リソース）を検索する検索システムにおいて、検索ＤＢを効率よく更新する。具体的には、検索される頻度が高いウェブページについては、他のウェブページよりもリンク切れのチェックを頻繁に行う。そのために、検索システムで構築されるキーワードＤＢを活用する。

【００２５】例えば、キーワードＤＢで使用回数が上位にランク付けされるキーワードを用いて検索ＤＢを検索した場合、検索結果としてＵＲＬの集合（以下、「キーワードにヒットする」という）が得られる。これらのＵＲＬは、頻繁に検索されるウェブページのＵＲＬと考えられる。そこで、このようなキーワードにヒットするウェブページについては、リンク切れのチェックを例えば１日１回行う。チェックの結果、リンク切れとなっているＵＲＬは検索ＤＢから削除される。

【００２６】上位のキーワードにヒットしないウェブページについては、従来通り例えば２週間に１回程度のチ

ェック頻度のままであるものの、ユーザにとってチェックの必要性が高いウェブページについて優先的にリンク切れをチェックするので、効率的に検索ＤＢを更新することが出来る。結果として、検索ＤＢから無効なＵＲＬが削除され、検索の速度及び精度が向上する。

【００２７】＜第１実施形態例＞

［構成］次に、本発明の検索システムについて、具体的に説明する。図１は、第１実施形態例に係る検索システムの全体構成図である。検索システムは、検索エンジン用サーバ（以下、単に検索サーバという）１１、ウェブサーバ１３及び利用者端末１４が、インターネット１５で接続されて構成されている。

【００２８】ウェブサーバ１３は、ウェブページを収集し、検索サーバ１１や利用者端末１４に提供している。利用者端末１４には、ブラウザ機能が搭載され、ウェブサーバ１３が提供するウェブページを表示可能である。また、利用者端末１４には、検索サーバ１１に検索要求を送信し、検索結果を出力するための検索クライアントが搭載されている。

【００２９】検索サーバ１１は、検索ＤＢ２１、キーワードＤＢ（ＫＷＤＢ）２２、ＤＢアクセス手段２３、検索要求処理手段２４、第１更新管理手段２５、第２更新手段２６、ＵＲＬ表２７及びリソースチェック手段２８を有している。検索ＤＢ２１には、ウェブページのＵＲＬが、キーワードに基づいて分類されて蓄積されている。検索ＤＢ２１には、希望者がＵＲＬを登録する登録型と、自動的にＵＲＬを収集して情報を構築するソフトウェアロボット収集型とがあり、そのどちらにも本発明を適用可能である。

【００３０】ＫＷＤＢ２２には、所定の順序でランク付けられたキーワードが蓄積されている。本実施形態例では、キーワードは、検索に使用された使用回数順にランク付けされている。図２は、ＫＷＤＢ２２の概念説明図である。ＫＷＤＢ２２には、検索キーワードと、そのキーワードが検索に使用された使用回数と、順位とが、記憶されている。本発明では、このＫＷＤＢ２２を利用してＫＷＤＢ２２の中で例えば上位１０位までについて、リンク切れなどのチェックを他のキーワードよりも短い時間間隔で行うようにする。例えば、通常は週１回の頻度でしかチェックされない場合には、上位のキーワードについては毎日１回チェックを行うようにする。

【００３１】検索要求処理手段２４は、利用者端末１４からの検索要求を受信し、キーワードにヒットするウェブページのＵＲＬの集合を、検索結果として利用者端末１４に送信する。ＤＢアクセス手段２３は、利用者端末１４からの検索要求に従って検索ＤＢ２１を検索する。また、ＤＢアクセス手段２３は、第１更新手段２５からの指示に従い、ＫＷＤＢ２２を更新する。さらにＤＢアクセス手段２３は、第２更新手段２６からの指示に従い、リンク切れしたＵＲＬを、検索ＤＢ２１から削除する。

【００３２】第１更新手段２５は、検索ＤＢ２１の検索に用いられたキーワードに基づいて、ＫＷＤＢ２２内のキーワードの順序を並べ替える。例えば、第１更新手段２５は、ＫＷＤＢ２２内のキーワードの新たな順位を計算し、キーワードの並べ替えをＤＢアクセス手段２３に指示する。第２更新手段２６は、リソースチェック手段２８を用いてインターネット上のウェブページを巡回し、リンク切れを検出する。また、第２更新手段２６は、ＫＷＤＢ２２で上位にあるキーワードを用いて検索ＤＢ２１を検索し、検索結果にヒットしたＵＲＬについてリソースチェック手段２８を用いてリンク切れしていないかどうかのチェックを行う。このチェックは、他のウェブページにおけるリンク切れの検出よりも高い頻度で行われる。また、第２更新手段２６は、重複するＵＲＬが検索結果に出現しないか否かのチェックも行う。これらのチェック結果に基づいて、第２更新手段２６は、検索ＤＢ２１の更新を行う。

【００３３】図３は、ＵＲＬ表２７の概念説明図である。検索ＤＢ２１には、キーワードに基づいて分類されたＵＲＬが蓄積されているが、その他にＵＲＬ表を設けておくと好ましい。ＵＲＬ表（ａ）では、チェック時間、検索キーワード、ＵＲＬ表（ｂ）へのポインタが記憶されている。「チェック時間」は、対応する検索キーワードにヒットするＵＲＬについてチェックした時間である。所定時間以上経過したエントリは、このチェック時間に基づいて削除される。「検索キーワード」は、検索に用いられたキーワードである。「ＵＲＬ表（ａ）へのポインタ」は、対応するキーワードで検索した場合にヒットするＵＲＬのリンク切れ検出回数が格納されているアドレスを示す。

【００３４】ＵＲＬ表（ｂ）では、リンク切れ検出回数と、ＵＲＬとが記憶されている。「リンク切れ検出回数」は、対応するＵＲＬについてリンク切れが検出された回数を示している。この例は、“ HYPERLINK http://www.fujitsu.co.jp www.fujitsu.co.jp”についてはリンク切れが検出されていないが、“wwww.fujitsu.co.jp”については３回のリンク切れが検出されていることを示している。

【００３５】リソースチェック手段２８は、検索サーバ１１から指示されたウェブページのデータを収集し、リンク切れが生じているか否かをチェックする。また、リソースチェック手段２８は、チェック結果を第２更新手段２６に通知する。

［処理方法］

（１）検出処理

図４は、検索システムが行う検出処理の流れの一例を示すフローチャートである。この処理では、検索システムは、ＫＷＤＢ２２で所定の順位のキーワードについて、リンク切れ及び重複ＵＲＬがあるか否かを、所定時間ご

とにチェックする。説明を容易にするために、１位のキーワードについて、１日１回チェックする場合を例に取る。

【００３６】ステップＳ１；第２更新手段２６は、ＫＷＤＢ２２で１位にランク付けされているキーワードを、ＤＢアクセス手段２３を介して取得する。

ステップＳ２；第２更新手段２６は、取得したキーワードを用いて検索ＤＢ２１の検索を行い、このキーワードにヒットするＵＲＬの集合を取得する。

ステップＳ３；第２更新手段２６は、取得したＵＲＬの集合の中から、重複するＵＲＬを削除する。

【００３７】ステップＳ４；第２更新手段２６は、取得した検索結果に基づいて、ＵＲＬ表を更新する。すなわち、検索に用いたキーワード及び取得したＵＲＬのためのエントリを、ＵＲＬ表（ａ）及び（ｂ）に設ける。既にＵＲＬがＵＲＬ表に登録されている場合は、そのＵＲＬについては新たなエントリを設けない。

ステップＳ５、Ｓ６；第２更新手段２６は、取得したＵＲＬについて、リンク切れが生じているか否かを、リソースチェック手段２８を用いてチェックする（Ｓ５）。リンク切れが生じていると、そのＵＲＬのリンク切れ検出回数をインクリメントする（Ｓ６）。

【００３８】ステップＳ７；第２更新手段２６は、検索したキーワードにヒットする全てのＵＲＬについて、リンク切れをチェックしたか否かを判断する。“Ｙｅｓ”と判断すると、ステップＳ８に移行する。“Ｎｏ”と判断すると、再び前記ステップＳ５に戻り、次のＵＲＬについてリンク切れをチェックする。

ステップＳ８；第２更新手段２６は、所定時間T1、本例では１日が経過するのを待機し、前記ステップＳ１に戻って前述の処理を繰り返す。

【００３９】本例では、ＫＷＤＢ２２で１位のキーワードについてのみ前述の処理を行ったが、例えば所定順位を１００位とする場合、各キーワードについて順次前述の処理を行えばよい。すなわち、第２更新手段２６は、１日に１回、ＫＷＤＢ２２で１位のキーワードにヒットするＵＲＬについて、リンク切れ及び重複ＵＲＬがあるかどうかをチェックし、その結果をＵＲＬ表（ａ）（ｂ）に書き込む処理を行う。ＫＷＤＢ２２で１位のキーワードにヒットしない他のウェブページについては、従来通り巡回して順次リンク切れをチェックする。後述するように、１位のキーワードについては、検索ＤＢ２１からリンク切れＵＲＬが削除されるので、検索速度及び精度を高めることが出来る。

【００４０】さらに、このＵＲＬ表に基づいて検索結果を出力すれば、検索結果から重複ＵＲＬが削除されているので、一層検索精度を向上することが出来る。この場合、第２更新手段２６は、所定時間T1よりも長い時間間隔T3でＵＲＬ表を参照し、チェック時間から所定時間が経過したエントリを削除する処理を行う。ＵＲＬ表は、上位のキーワードによる検索結果を示しているものの、あまりにも古い検索結果を用いることを防止するためである。

【００４１】（２）ＤＢ更新処理

図５は、第２更新手段２６が行うＤＢ更新処理の流れを示すフローチャートである。この処理では、所定時間T2（T2＞T1）毎に、リンク切れが所定回数以上のＵＲＬをＵＲＬ表及び検索ＤＢ２１から削除する処理を行う。

ステップＳ２１；第２更新手段２６は、所定時間T2が経過するのを待機し、経過するとステップＳ２２に移行する。

【００４２】ステップＳ２２、Ｓ２３，Ｓ２４；第２更新手段２６は、その時点で作成されているＵＲＬ表（ｂ）を参照し、リンク切れ検出回数が所定数以上のＵＲＬを検索する。そのようなＵＲＬが存在する場合、ＵＲＬ表（ａ）（ｂ）及び検索ＤＢ２１から該当ＵＲＬを削除する。

ステップＳ２５；第２更新手段２６は、ＵＲＬ表（ｂ）の全てのエントリについて、リンク切れ検出回数をチェックしたか否かを判断す。“Ｙｅｓ”と判断すると前記ステップＳ２１に戻って前述の処理を繰り返す。“Ｎｏ”と判断すると、ＵＲＬ表（ｂ）の次のエントリのＵＲＬについて、リンク切れ検出回数をチェックする（Ｓ２２）。

【００４３】すなわち、この処理では、リンク切れ検出回数が所定数以上のＵＲＬを、ＵＲＬ表（ａ）（ｂ）から削除し、またそのＵＲＬを検索ＤＢ２１から削除する。これにより、無用なＵＲＬが検索ＤＢ２１から削除されて検索処理速度が上がることを期待できる。さらに、無用なＵＲＬが検索結果に出現しなくなり、検索精度を高めることができる。

【００４４】なお、リンク切れが１回でもあればＵＲＬ表及び検索ＤＢ２１から削除することも考えられる。しかし、諸原因、例えばネットワークのトラフィックやウェブサーバの一時的なトラブルによりリンク切れではないのにアクセスできない場合を考慮し、若干の余裕を持たせて削除を行うことが好ましい。

＜第２実施形態例＞図６は、第２実施形態例に係る検索システムの全体構成図である。この検索システムは、ＫＷＤＢ２２が複数設けられている点を除き、前記第１実施形態例と同様の構成を有する。

【００４５】この例では、複数のＫＷＤＢ２２、すなわちＫＷＤＢ-ａ及びＫＷＤＢ-bが、ドメイン名毎に設けられている。具体的には、ＫＷＤＢ-aは、利用者端末１４のドメイン名が“co.jp”の場合に用いられる。ＫＷＤＢ-bは、利用者端末１４のドメイン名が“ne.jp”の場合に用いられる。例えば、会社から仕事に関するウェブページを検索する場合と、自宅から趣味に関するウェブページを検索する場合とを考える。仕事に関するウェブページと趣味に関するウェブページとでは、内容やア

クセス頻度が異なる場合が多い。従って、同じＫＷＤＢ２２を使用するのは適切ではない場合がある。

【００４６】例えば、仕事に関する分野では、キーワードがＫＷＤＢ２２の上位にランキングされ、ＵＲＬのチェックが頻繁に行われる。一方、趣味に関する分野では、その分野ではよく使用されるキーワードであっても全体としてのキーワードの使用回数が少ないため、ＫＷＤＢ２２の上位にランク付けされない。結果として、この分野のウェブページについて、リンク切れチェックの頻度が低下する。

【００４７】そこで、所定の分類、例えば利用者端末１４のドメイン名毎にＫＷＤＢ２２を設け、検索要求を行った利用者端末１４のドメイン名に応じて使用するＫＷＤＢ２２を切り替える。利用者端末１４のドメイン名は、インターネットで用いられるhttpプロトコルに準拠して検索サーバ１１に送信されるＩＰアドレスに通常含まれているので、そこから取得可能である。

【００４８】図7は、ドメイン名毎にＫＷＤＢ２２を更新する処理の流れを示すフローチャートである。この処理では、検索要求がある毎に、ドメイン名に応じてＫＷＤＢ２２を更新する処理を行う。

ステップＳ３１；検索要求処理手段２４は、利用者端末１４から検索要求を受信する。

【００４９】ステップＳ３２；検索要求処理手段２４は、利用者端末１４のＩＰアドレスなどからドメイン名を抽出する。

ステップＳ３３；検索要求処理手段２４は、利用者端末１４のドメイン名に対応するＫＷＤＢ２２を選択する。

ステップＳ３４；検索要求処理手段２４は、検索要求に含まれる検索キーワードを抽出する。

【００５０】ステップＳ３５；検索要求処理手段２４は、ＤＢアクセス手段２３を用い、選択したＫＷＤＢ２２に抽出した検索キーワードが含まれているか否かを判断する。含まれていれば、そのキーワードの使用回数をインクリメントとし、必要に応じて順位を変更する。含まれていなければ、そのキーワードをＫＷＤＢ２２に登録し、使用回数を“１”とする。

【００５１】ステップＳ３６；検索要求処理手段２４は、検索キーワードに基づいて検索ＤＢ２１の検索を実行し、検索結果を利用者端末１４に送信する。

＜第３実施形態例＞前記第１及び第２実施形態例では、ＫＷＤＢ２２においてキーワードの使用回数順にキーワードをランク付けし、使用回数が多いキーワードについて優先的に検出処理及びＤＢ更新処理を行っている。しかし、他の順序でキーワードをランク付けすることも考えられる。本実施形態例では、所定期間内における使用回数の変化に基づいて、優先的に検出処理及びＤＢ更新処理を行う。

【００５２】図８は、本実施形態例におけるＫＷＤＢ２２の概念説明図である。このＫＷＤＢ２２には、検索キーワード、使用回数、順位、一定時間前の順位、変化速度、一定時間前の変化速度及び変化加速度が記録されている。このＫＷＤＢ２２の順位は、一定時間、この例では一週間毎に再計算される。「順位」は、変化加速度の大きさに基づいて決定されている。「一定時間前の順位」とは、例えば一週間前の順位である。「変化速度」とは、「一定時間前の順位」から「順位」への変化を示す。「一定時間前の変化速度」とは、例えば一週間前に置ける変化速度である。「変化加速度」とは、「一定時間前の変化速度」から「変化速度」への変化率を示す。具体的には、先々週の変化速度から先週の変化速度への変化率を示す。

【００５３】図８では、キーワード「富士通」の順位が１位になっている。このキーワードの一週間前の順位は２位である。先週の順位と今週の順位との差、すなわち「変化速度」は「１」となる。一週間前の時点における変化速度、すなわち「一定時間前の変化速度」は１０となっている。先週と先々週との変化速度の差、すなわち「変化加速度」は「－９」となる。

【００５４】このように、所定期間内における使用回数の変化に基づいてキーワードの順位を決定すると、例えば急激に関心が高まった分野のウェブページについてリンク切れなどのチェックをすることができる。従って、その時々のニーズに応じた検索ＤＢ２１の更新を行うことが可能となる。

＜第４実施形態例＞前記第３実施形態例においては、所定期間内における使用回数の変化に基づいて、ＫＷＤＢ２２内のキーワードの順位を決定している。本実施形態例では、使用回数の変化に加え、さらに前記図８の各項目に重み付けを行ってキーワードをランク付けする。

【００５５】例えば、重み付けパラメータとして、「順位」に対し０．５、「変化速度」に対し０．３を用いる。これらの項目に重み付けパラメータを乗じた値に基づいて順位を算出する。具体的には、次のようにして算出する。図８のキーワード「富士通」の順位は１位であり、一定時間前の順位は２位である。先週からの順位の変化を示す変化速度は、２－１＊０．５＝１．５と計算される。また、変化加速度、すなわち先週と先々週との変化速度の差は、１．５－１０＝－８．５と算出される。このようにして算出した変化加速度の値を用いてキーワードのランク付けを行う。

【００５６】＜第５実施形態例＞第５実施形態例においては、放送禁止用語、差別用語などに相当する言葉にヒットするウェブページの変化を検索ＤＢ２１に反映させない。まず、放送禁止用語、差別用語などに相当する言葉を、禁止キーワードとして検索ＤＢ２１にあらかじめ登録しておく。検索サーバ１１は、ＫＷＤＢ２２で上位のキーワードが禁止キーワードに該当するか否かの判断を行い、禁止キーワードに該当する場合、検索ＤＢ２１の更新処理を行わない。

【００５７】図９は、検索サーバ１１が行う処理の流れを示すフローチャートである。
ステップＳ４１；検索要求処理手段２４は、検索要求を利用者端末１４から受信する。
ステップＳ４２；検索要求処理手段２４は、検索要求に含まれる検索キーワードを抽出し、第１更新手段２５に通知する。
【００５８】ステップＳ４３；第１更新手段２５は、抽出した検索キーワードが所定の禁止キーワードに含まれているか否かを判断する。“Ｙｅｓ”と判断すると後述するステップＳ４５に移行し、そのまま通常の検索処理を実行する。“Ｎｏ”と判断すると、ステップＳ４４に移行する。
ステップＳ４４；第１更新手段２５は、検索キーワードに基づいて、ＫＷＤＢ２２を更新する。まず、抽出した検索キーワードがＫＷＤＢ２２に含まれているか否かを判断する。含まれていれば、そのキーワードの使用回数をインクリメントとし、必要に応じて順位を変更する。ＫＷＤＢ２２に含まれていなければ、そのキーワードをＫＷＤＢ２２に新たに登録し、使用回数を“１”とする。
【００５９】ステップＳ４５；検索サーバ１１は、検索キーワードに基づいて検索ＤＢ２１を検索し、検索結果であるＵＲＬの集合を利用者端末１４に送信する。この実施形態例を用いれば、公序良俗に反する内容のウェブページについてまで頻繁にリンク切れをチェックすることがなくなり、無用なチェックによる負担を省略でき、社会の道徳に適合した検索ＤＢ２１の更新を行うことが出来る。
【００６０】＜第６実施形態例＞第６実施形態例では、ＫＷＤＢ２２の構築にあたり、あらかじめ語句の類似関係を定義する類似用語辞書を検索ＤＢ２１に登録しておく。ＫＷＤＢ２２を更新する場合、類似範囲に属するキーワードは同一のキーワードとして扱う。図１０は、類似用語辞書の概念説明図である。この図では、漢字、ローマ字、ひらがな、カタカナによる差しかない語句を、類似と見なす辞書の例を示している。類似用語辞書としては、通常用いられる辞書を使用できる。
【００６１】本実施形態例に依れば、ひらがな、カタカナあるいはローマ字による違い、聴音や空白の有無の違いなどのみであり、実質的な違いのないキーワードをまとめて同一の語句としてＫＷＤＢ２２で扱うことが出来る。従って、ＫＷＤＢ２２のランク付けが実情に即したものとなり、さらには検索ＤＢ２１の更新が実情に即したものとなる。
【００６２】＜第７実施形態例＞前記第２実施形態例においては、複数のＫＷＤＢ２２をドメイン名毎に構築している。しかし、他の分類に基づいて複数のＫＷＤＢ２２を構築することも可能である。例えば、個人情報を蓄積したユーザＤＢが検索システム上にあらかじめ存在している場合、個人情報に基づいてＫＷＤＢ２２を構築しても良い。個人情報とは、例えばユーザ名、性別、居住地域、年代、趣味などである。
【００６３】例えば、年代に応じたＫＷＤＢ２２を構築しておき、検索要求があった場合、要求元ユーザの年代に対応するＫＷＤＢ２２を更新する。これにより、ＫＷＤＢ２２のランク付けが各ユーザの年代に適合した内容となり、検索ＤＢ２１は各年代のユーザの要求に合うように更新される。ＫＷＤＢ２２を構築する他の分類として、アクセス時間毎にＫＷＤＢ２２を構築しておくことも考えられる。例えば、午前中はニュース関係、午後は株式関係、夜は趣味関係の検索が多いとする。その場合、午前中、午後、夜用のＫＷＤＢ２２を準備しておく。各ＫＷＤＢ２２は、そのＫＷＤＢ２２の時間帯における検索要求に応じて更新される。
【００６４】このようにすると、各ＫＷＤＢ２２におけるキーワードの順位は、時間帯に応じた内容を反映したランク付けとなる。また、各自時間帯でアクセスの多い分野において、無効なＵＲＬを検索ＤＢ２１から取り除くので、検索ＤＢ２１を各時間帯における利用者のニーズに適合させることが出来る。
＜他の実施形態例＞
（Ａ）検索システムを複数のサーバで構成することも可能である。図１１は、複数のサーバで構成された検索システムの構成例を示す。この図に示す検索システムは、前記検索サーバ１１に相当する部分が、サーバ群Ａと、リソースチェック手段を搭載したサーバ群Ｂと、検索ＤＢ２１と、ＫＷＤＢ２２と、ＵＲＬ表２７とから構成されている。
【００６５】サーバ群Ａを構成するサーバには、ＤＢアクセス手段２３、検索要求処理手段２４、第１更新手段２５及び第２更新手段２６が設けられている。また、サーバ群Ｂを構成するサーバには、リソースチェック手段２８が設けられている。これらの機能は、前記第１実施形態例において述べたものと同様である。検索ＤＢ２１、ＫＷＤＢ２２及びＵＲＬ表２７には、前述の実施形態例で例示したような情報が蓄積される。これらは、それぞれ独立に存在してサーバ群Ａ、Ｂから直接アクセスされたり、データベースサーバを介してサーバ群Ａ、Ｂからアクセスされたりする。
【００６６】（Ｂ）前述した本発明の方法を実行するプログラムを記録した記録媒体は、本発明に含まれる。ここで記録媒体としては、コンピュータが読み書き可能なフロッピー（登録商標）ディスク、ハードディスク、半導体メモリ、CD-ROM、ＤＶＤ、光磁気ディスク（ＭＯ）、その他のものが挙げられる。
【００６７】
【発明の効果】本発明を用いれば、利用者の関心が高いと考えられる分野において優先的にリソースをチェックし、無駄になったリソースを検索蓄積手段から削除する

ことができる。従って、検索速度及び検索精度の向上を、利用者の関心が高い分野において期待することが出来る。

【図面の簡単な説明】

【図１】第１実施形態例に係る検索システムの構成図。

【図２】ＫＷＤＢの概念説明図。

【図３】ＵＲＬ表の概念説明図。

【図４】検出処理の流れを示すフローチャート。

【図５】ＤＢ更新処理の流れを示すフローチャート。

【図６】第２実施形態例に係る検索システムの構成図。

【図７】複数ＫＷＤＢの更新処理の流れを示すフローチャート。

【図８】第３実施形態例におけるＫＷＤＢの概念説明図。

【図９】禁止キーワードチェック処理の流れを示すフローチャート。

【図１０】類似キーワード辞書の概念説明図。

【図１１】複数のサーバで構成された検索システムの構成図。

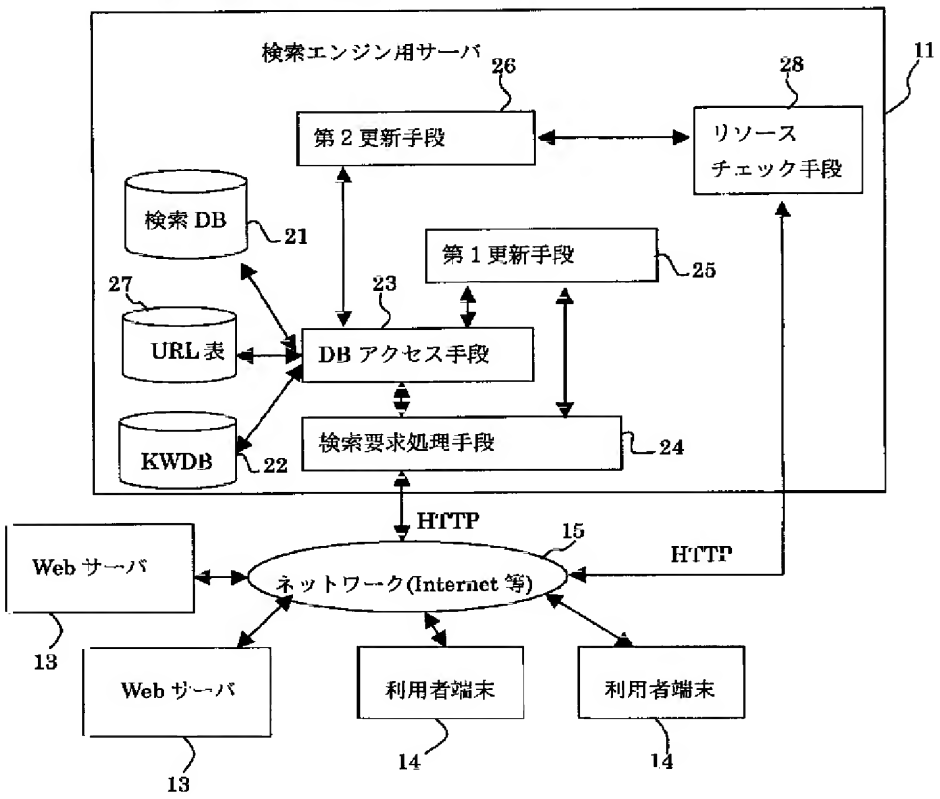
【符号の説明】

２１；検索ＤＢ

２２；ＫＷＤＢ

【図１】

第１実施形態に係る検索システムの構成



【図２】

ＫＷＤＢの概念説明図

検索キーワード	使用回数	順位
富士通	123456	1
研究所	2345	2
…	…	…

【図3】

URL表の概念説明図

(a)

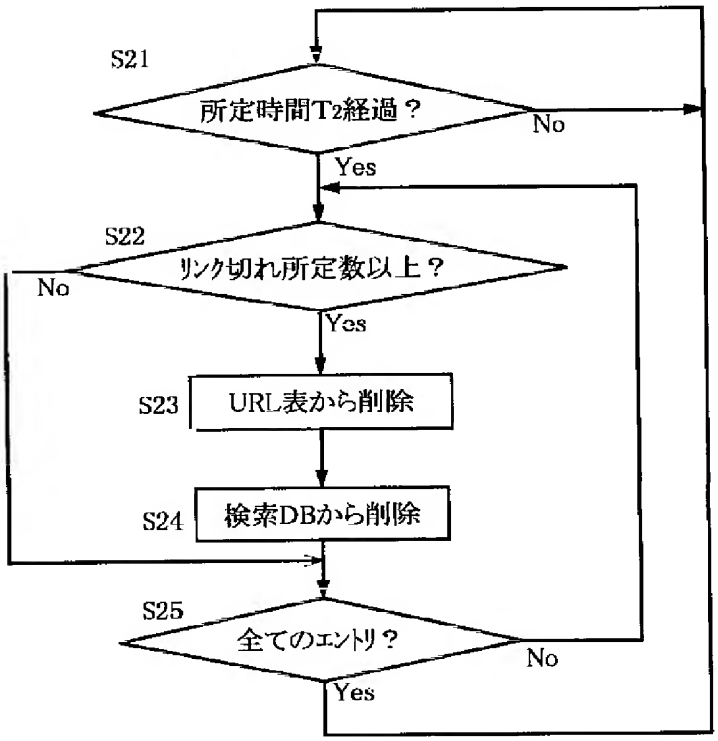
チェック時間	検索キーワード	URL 表へのポインタ
2000/4/4/15:20	富士通	*** **
2000/4/4/15:25	研究所	*****

(b)

リンク切れ検出回数	URL
0	www.fujitsu.co.jp
3	www.fujitsu.co.jp

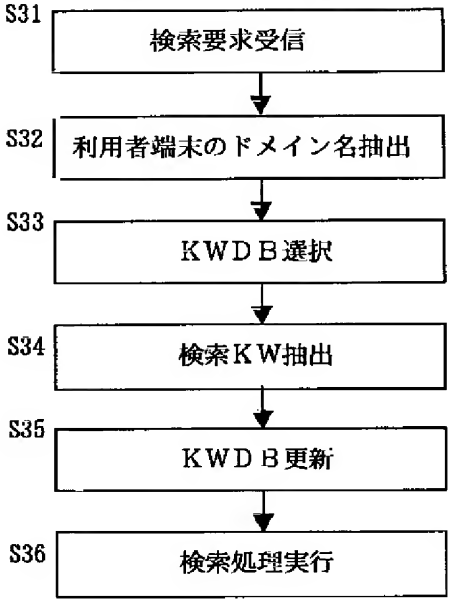
【図5】

DB更新処理の流れ



【図7】

複数KWDBの更新処理



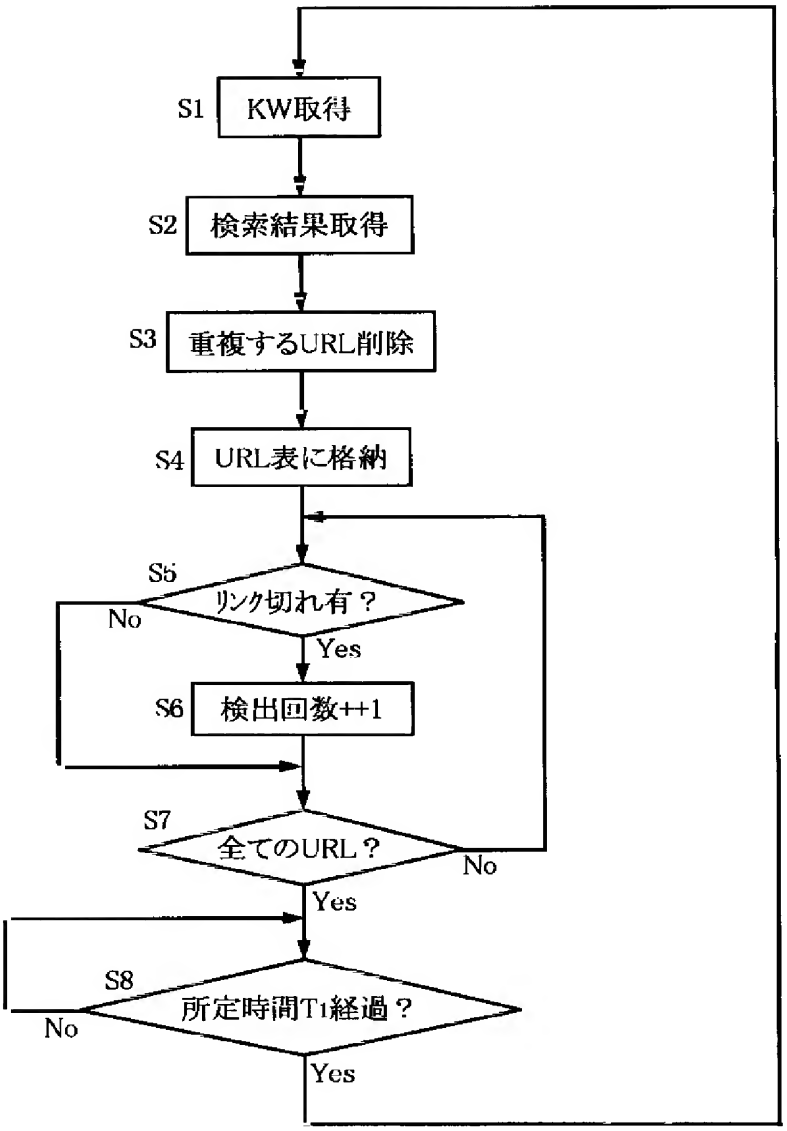
【図8】

第3実施形態例におけるKWDBの概念説明図

検索キーワード	使用回数	順位	一定時間前の順位	変化速度	一定時間前の変化速度	変化加速度
富士通	12345	1	2	1	10	-9
研究所	2345	2	12	10	5	5
…	…					…

【図４】

検出処理の流れ



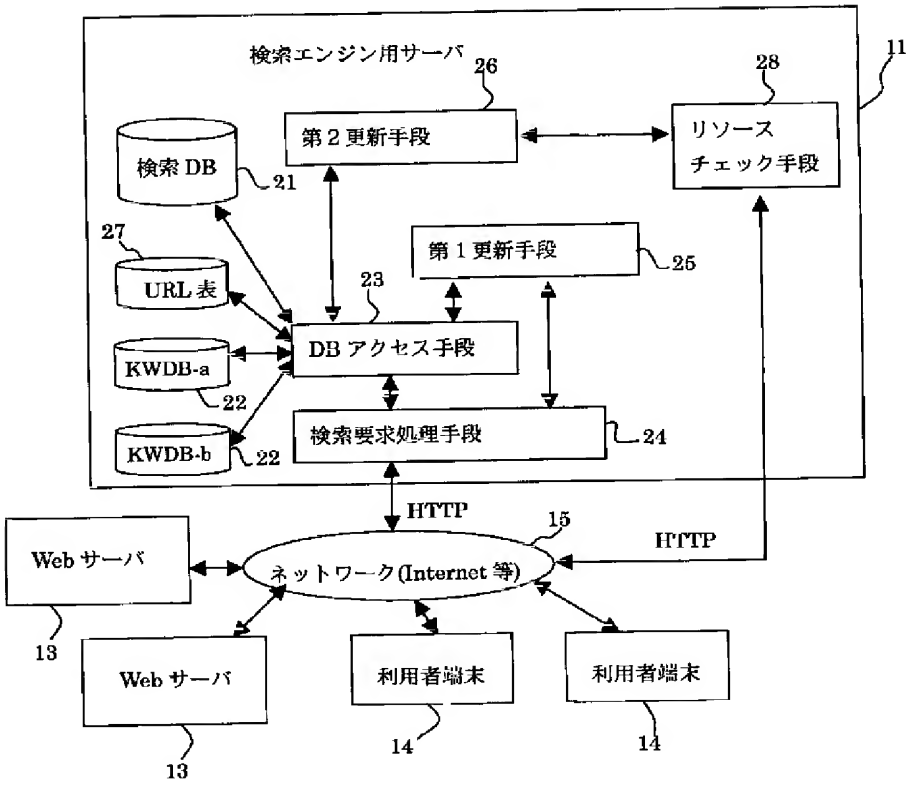
【図１０】

類似キーワード辞書の概念説明図

検索キーワード	類似候補	類似候補	類似候補
富士通	FUJITSU	ふじつう	フジツウ
研究所	Laboratories	Laboratory	ラボ
…			…

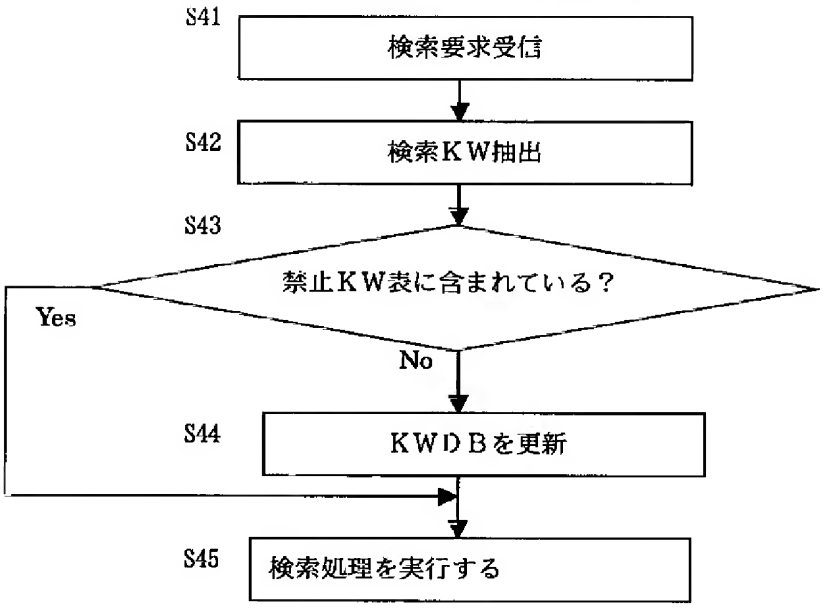
【図６】

第２実施形態に係る検索システムの構成

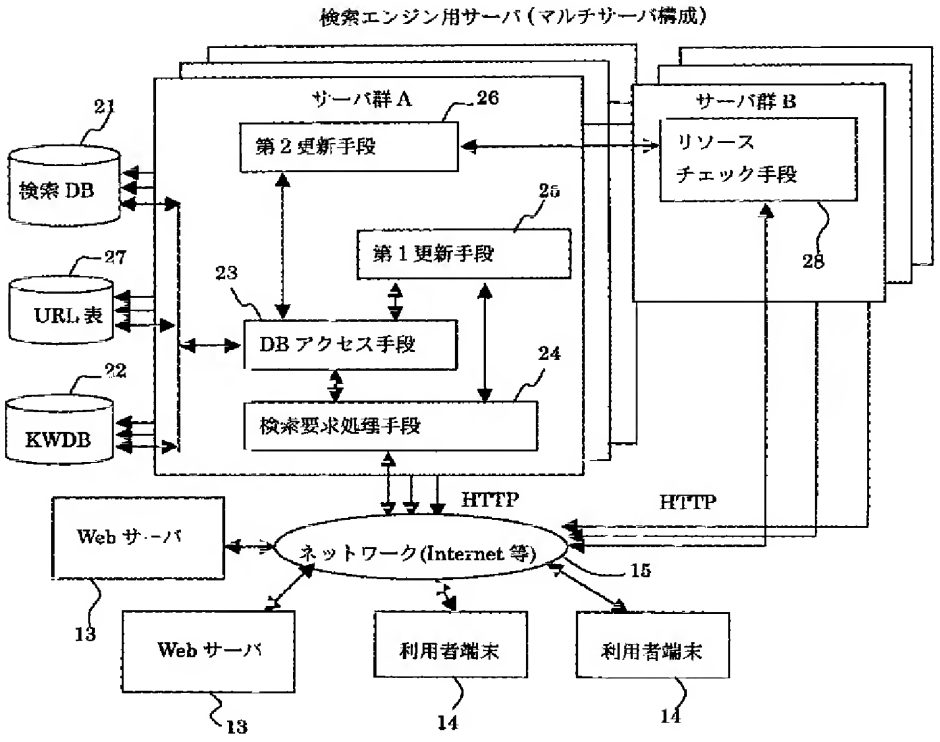


【図９】

禁止キーワードチェック処理の流れ



【図１１】



フロントページの続き

(51)Int.Cl.[7]	識別記号	ＦＩ		(参考)
Ｇ０６Ｆ　13/00	５１０	Ｇ０６Ｆ　13/00	５１０Ｂ	

(72)発明者　木原　英人
神奈川県川崎市中原区上小田中４丁目１番１号　富士通株式会社内

Ｆターム(参考)　5B075 NK02 NR05 PR04
5B082 HA00